Frequency-Domain Auditory Suppression Modelling (FASM) – A WDFT-Based Anthropomorphic Noise-Robust Feature Extraction Algorithm For Speech Recognition

Alexei V. Ivanov*, Marek Parfieniuk**, and Alexander A. Petrovsky***

* - Speech Technology Center, Moscow, Russian Federation
** - Real Time Systems Department, Faculty of Computer Science, Bialystok Technical University, Bialystok, Poland
*** - marekpk@ii.pb.bialystok.pl, *** - palex@it.org.by

Abstract

This paper presents a physiologically inspired feature extraction algorithm for employment within the speech recognition engines, which are supposed to remain effective in noisy environments. Essentially, the algorithm simulates a key property of the “active cochlea” models – a signal dependent variable gain over the frequency range. In order to drastically reduce computational complexity of the algorithm in comparison to the original time domain “active cochlea” models, it is implemented in the frequency domain with the help of a warped discrete Fourier transformation (WDFT). The essence of FASM technique is that in the presence of the noise, higher frequency channels get more attenuation if there are “enough” signal components in the lower, less susceptible to the noise influence, part of the spectrum. As it is confirmed by the performed measurements FASM algorithm allows to boost feature invariance to noise while keeping feature informativeness at the acceptable level.

1. Introduction

The anthropomorphic, i.e. acting in the same way as human organs, signal processing algorithms represent one of the most promising approaches to robustify the existing automated speech recognition (ASR) devices towards different types of interference, which is inherent to a realistic acoustic environment. The performance comparison in various speech recognition tasks in the realistic environments shows clear advantage of humans over machines, e.g. while machines are heavily impaired by the noise with SNR=0db, humans have only a minor, if any, discomfort performing at the same level in these conditions. Particularly, experiments with recognition of random digit sequences in adverse environments follow the depicted pattern. This observation supports a conclusion that neither improvements in acoustic nor language modelling but enhancement of the feature extraction may solve the robustness problem. The unreasonable high susceptibility of the existing ASR to the environment interference is perceived as one of the major obstacles to wide application of ASR technology. It seems obvious that the algorithmical interpretation of human audio signal processing and employment of the obtained algorithms in ASR systems may indeed bridge the mentioned performance gap. This approach constitutes to the essence of the anthropomorphic modelling.

Majority of the existing feature extraction algorithms generally follow the same signal processing path as the live systems. The first step in generation of the feature vector is to obtain a perceptual-like spectrogram of the observed speech utterance in the time-frequency plane. This step is in agreement with the “spectral analysis” of the cochlea. The STFT-based sonogram acquisition assumes, that after calculation of a uniform-resolution signal spectrum certain adjacent bins should be integrated in accordance with one of the perceptual frequency scales. The depicted way of calculation is thought to have certain computational redundancy, since the calculated frequency bins are only used for perceptual integration, thus spectral resolution becomes wasted.

The psychoacoustical auditory modelling contains a controversy. While admitting that human auditory apparatus is a non-linear signal processing tool, the psychoacoustical approach attempts to infer its properties in the experiments with simple sounds (e.g. single tones, noises, clicks) and generalises the observed relation to the case of such complex non-stationary signal as speech. One of the possible alternatives to the psychoacoustical modelling, which sets up a goal of empirical characterisation of listener’s subjective response to a given sound, is a physiological modelling approach, which divides the “non-linear black box” of auditory apparatus into a number of elements and concentrates on the description of the objectively measured processes involving them.

Besides already noted ability to translate frequency of the applied excitation into a place of maximal activity along the basilar membrane, the inner ear is known to exhibit a property of auditory suppression, i.e. the objective inhibition of the basilar membrane response to a certain signal component in the presence of the other signal components in the lower part of the spectrum. Auditory suppression along with a “spread of excitation” phenomenon are viewed as the main mechanisms of the psychoacoustically observed masking. A class of “active cochlea” models (e.g. given in [1,2]) is capable of capturing and explaining of the suppressive cochlear action.

Among the main reasons why the suppression modelling is supposed to give robustness advantage to ASR are the observations that the suppressive action boosts contrast of the “perceptual spectrogram” in perception of vowels [3] and additionally the fact, that patients with auditory pathologies,
linked to the action of cochlear amplifier, experience difficulties in speech recognition in noisy environments [4]. The present paper is concentrated on the further development of the idea of the auditory suppression modelling application [1,2] to increase robustness of the ASR feature extraction towards the adverse acoustic interference. Here we attempt to depart from very detailed model of the cochlear amplifier [1] and develop a simplified and computationally effective frequency-domain suppression model and assess it’s merits to the ASR feature extraction.

2. Frequency-domain auditory suppression modelling (FASM)

Utilisation of the Warped Discrete Fourier Transform (WDFT) [5] gives an advantage of the direct transformation of incoming signals to their spectral representation, with the frequency scale being warped to be perceptual-like. A specific frequency warping is defined with a conformal map in the \( z \)-plane, which transforms the unit circle into itself. Particularly, a map, which is defined by the first order all-pass filter of the form:

\[
A(z) = \frac{\rho + z^{-1}}{\rho z + 1}
\]

with \( 0 < \rho < 1 \), allows to stretch spectral bins in the low-frequency range and compress them in the high-frequency range:

\[
D\left(\frac{\phi}{2}\right) = \frac{1 - \rho}{1 + \rho} \left(\frac{\gamma}{2}\right)
\]

here \( \gamma \in [0, \pi] \) is a “linear” frequency scale (which constitutes to the unit circle in the \( z' \)-plane), and \( \phi(y) \in [0, \pi] \) is a “warped” frequency scale in the \( z' \)-plane by transformation (1). Frequencies \( \phi \) and \( \gamma \) are normalised and thus do not depend upon sampling frequency \( F_s \).

The physiological frequency scale is defined by an empirical formula, which approximates the frequency-place transformation of the human cochlea [6]:

\[
f(x) = 160 \left( 10^{2.1x} - 0.8 \right)
\]

here \( f(x) \in [F_{min}, F_{max}] \) is a frequency (expressed in Hz), which constitutes to a certain location \( x \) on the basilar membrane, \( x \in [0,1] \) is a normalised distance along the basilar membrane from a chosen location to the cochlea apex. From the practical point of view it makes sense to simplify the relation (3) by substituting the coefficient 0.8 with 1. This change leads to a minor modification of the frequencies, corresponding to the cochlea locations \( x > 0 \), and makes exact correspondence \( f(0) = 0 \), that is convenient assuming the task of approximation of the scale (3) by the scale (2). Con-version of (3) to normalised frequencies gives the following expression:

\[
\psi = 320 \pi \left( \frac{F_s^2}{320} + 1 \right)^{-1} \cdot \left( F_s + 1 \right)^{-1}
\]

where \( \gamma \in [0, \pi] \) is the same “linear” frequency scale that is used in (2) and \( \psi(y) \in [0, \pi] \) is an ideal physiological frequency scale, which approximation with the scale \( \phi(y) \) is desired.

Comparison of the relations (2) and (4) makes it apparent that the choice of \( \rho \) and \( F_s \) governs a degree of similarity between these two frequency scales. The performed numerical experiment shows that a value of \( \rho = 0.60071382564373 \) minimises the mean squared difference between \( \phi(y) \) and \( \psi(y) \):

\[
S(\psi, \phi) = \int_0^\pi \left( \psi(y) - \phi(y) \right)^2 dy
\]

with an \( y \) scale step being equal to \( 0.001 \pi \) and an a priori defined value of \( F_s = 16\text{kHz} \).

Warped DFT \( X(k), k \in \mathbb{N}, 0 \leq k < N \) is being calculated by multiplication of the incoming signal frame \( x(k), k \in \mathbb{N}, 0 \leq k < N \) by DFT matrix in \( z' \)-domain, that is expressed through \( z' \)-domain variables with the help of substitution (1) [7]. It is possible to assemble separate WDFT representations of the input signal frames into perceptual spectrogram in the similar manner as the classical STFT. Further stages of FASM algorithm (fig.1) include computation of the power spectrum \( P(k), k \in \mathbb{N}, 0 \leq k < N / 2 + 1 \), estimation of a signal-dependent sub band gain \( G(k) \), weighting of the power spectrum with the estimates of the \( G(k) \); calculation of the LP-cepstral coefficients and estimation of their first and second derivatives.

![Figure 1: FASM feature extraction structure.](image-url)

Further throughout the paper the term “baseline” algorithm refers to an essentially identical way to obtain speech features, but without the step of the signal-dependent sub band gain \( G(k) \) estimation.

The values of a gain parameter \( G(k) \) in each of the perceptual sub bands are obtained by multiplication of the \( G \) matrix by the vector of the power spectrum \( P(k) \). The matrix \( G \) is defined as a sum of matrix, which contains in its rows...
gains of the \( G \) filters of the active cochlea model [1,2] at the central frequencies of each of the perceptual sub bands, and the unity matrix. The product is further subjected to a compressive non-linearity, which is implemented as a sigmoid. Thus the gain parameters are calculated in accordance with the following formula:

\[
G(k) = 1 - \frac{(G_{\text{max}}(k) - G_{\text{min}}(k))}{2} \cdot \left(1 + \exp\left(-\frac{1}{C} \left(P(k) + \sum_{l} P(l) \cdot G_{l}(l)\right)\right)\right) - 1
\]  

(6)

which guaranties that each sub band gain changes between its maximal (unity) and minimal \( (G_{\text{max}}(k)/G_{\text{min}}(k)) \) values depending on the total signal power in the lower than the chosen perceptual domains. It also reflects the effects of the \( G_i \) filters. The values \( G_{\text{max}}(k) \) and \( G_{\text{min}}(k) \) stand for the minimal and maximal possible gains of the active cochlear model [1,2] and define here the appropriate gain dynamic range in each of the perceptual sub bands. The coefficient \( C \) is a scaling factor, which defines the amount of suppressive action of a given input signal level. In the present work it was chosen to be \( 10^4 \).

The depicted suppression model has a number of simplifications compared to the original active cochlea model. In the first place, \( P(k) \) multiplication coefficient in (6) must be \( G(k,t-1) \), not unity. Due to the fact that the WQFT analysis frame spans a considerable amount of time it was decided to avoid modelling of the gain transitional processes. Secondly, it was assumed that the amount of suppression depends upon the total signal power. This assumption does not precisely capture the exact nature of the phenomena in the inner ear.

3. Application of FASM in terms of the missing feature theory

In [8] it is explicitly pointed, that “masked data is effectively missing data” and the whole missing feature theory (MFT) is inspired by the role of masking in the process of sound perception by humans. While extensively exploited in the field of speech coding models of masking did not find general acceptance so far in ASR field. From the first glance, application of any masking model will lead to the loss of information in the incoming signal. However, spectral integration techniques (like MFCC or PLP) are also aimed at reducing amount of information needed to describe the incoming spectral pattern in order to alleviate the “curse of dimensionality” problem during acoustic model training. The key here is to infer which part of the total information being brought by the acoustic signal characterises the speech source and retain as much of this information as possible. This reasoning leads to a conclusion that feature extraction algorithms must be built in accordance with the maximal mutual information (MMI) criterion.

Missing feature theory concentrates on the proper handling of situations when one has apriory knowledge of which part of the source characterisation was most probably altered by the communication channel and, thus, deemed to be unreliable in the source state sequence decoding process. MFT does not explicitly specify the procedure to classify signal components into “reliable” and “unreliable” classes. Variety of different techniques was proposed to perform this classification: starting from the most obvious “low-SNR” thresholding to perceptually inspired masking models.

Employment of suppression modelling instead of masking models is more theoretically justified. As the auditory apparatus is a non-linear time-variant system, the subjective psychoacoustical characterisation that certain signal component is “finally not heard” by the whole auditory system does not guarantee that it did not play it’s role at the early auditory processing stages. If we consider comparison of the FASM output spectrum with a certain level, the proposed algorithm may serve as a mean to perform the “reliable/unreliable” labelling. Unlike plain STFT in “low-SNR” approach, activity in any particular FASM sonogram point depends not only upon the strength of the corresponding signal component, but also on the instantaneous signal-dependent susceptibility of the model to that component. Thus certain signal components might be inhibited under the influence of other parts of incoming signal, even though originally they were quite intense. To be good in “reliable/unreliable” labelling FASM shall inhibit those signal components, which are dominated by channel rather than source and consequently make output more invariant to channel variation.

4. FASM algorithm performance measurements

The measurements of the algorithm performance include estimation of the mutual information between a speech source and the features complemented with assessment of the degree of feature invariance towards additive noises at the various levels. Although it is widely perceived that a speech recognition experiment serves as a final measurement of the technique viability, we stick to the performance assessment in accordance with the criterions formulated above. Indeed, speech recogniser is typically a complex device with many estimated parameters, which affect the overall performance and, thus, potentially there are many things which may go wrong and distort the performance picture.

The mutual information was measured on the basis of phonetically labelled speech from TIMIT database. The exact procedure of this experiment is described in [9]. Essentially it consists of a non-parametric probability distribution histogram estimation, which is further used for the mutual information calculation. Amount of the histogram bins that is needed for reliable estimation of the true distributions was taken after Akaike information criterion.

Results of this experiment are given in the fig. 2. The main conclusion, which can be maid from the results is that FASM suppression modelling only slightly reduces informative content of the feature vector. It happens mainly due to the diminished informativeness of the cepstral coefficient \( c_0 \) (feature no.1, fig.2), which is to the large extent compensated for by an increased informativeness of the higher order cepstral coefficients. Inclusion of the “zero” cepstral coefficient into the feature vector is a subject to a trade off between the considerable boost in a cumulative (not total, because the cepstral features are proved to be mutually dependent for speech signals) informativeness of the feature vector and unwanted dependency of the feature vector upon the incoming speech message loudness. It is still possible to
approximately reconstruct the signal power spectrum on the basis of the higher order LP-cepstral coefficients, but in such a way that it’s total power is always kept constant. Thus, exclusion of the \( c_k \) from the feature vector may be viewed as a normalisation of the spectrogram in such a way, that the intense signal components reduce susceptibility to fainter ones regardless to their spectral position. This normalisation may also be viewed as a very rough model of masking. Obviously, the depicted “cepstral masking model” is not capable of increasing feature informativeness. Consequently, we may conclude, that the property of suppression to spread from the lower part of the spectrum to the higher one is important, because it allows to increase informativeness of the certain higher order cepstral coefficients.

**Figure 2: Feature informativeness.**

Procedure of the invariance measurement is also described in [9]. Estimation of the Euclidian distance between the clean and noisy features, as a measure of the feature invariance, was accomplished on TI20 recordings with various amounts of the additive white noise.

**Table 1: Feature invariance to noise**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>SNR=0db</th>
<th>SNR=10 db</th>
<th>SNR=20 db</th>
<th>SNR=30 db</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.6520</td>
<td>0.5198</td>
<td>0.4479</td>
<td>0.2873</td>
</tr>
<tr>
<td>Cepstrum N1-9</td>
<td>0.5470</td>
<td>0.4609</td>
<td>0.4091</td>
<td>0.2753</td>
</tr>
<tr>
<td>Cepstrum N2-9</td>
<td>0.1050</td>
<td>0.0589</td>
<td>0.0387</td>
<td>0.0120</td>
</tr>
<tr>
<td>Difference</td>
<td>0.1050</td>
<td>0.0589</td>
<td>0.0387</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

The experiment results, which are given in table 1, show that FASM algorithm generates more invariant features than the baseline either in the absence (upper part of the table) or presence (lower part of the table) of zero cepstral coefficient. In the case of intense noise (SNR ~ 0-10db) FASM Euclidian distances are almost the same as “baseline” distances for a 10db less noisy cases. It is also evident, that the positive effect of FASM grows in amplitude as noise becomes stronger. Estimation of the confidence interval for this experiment (procedure is depicted in [9]) confirms that the positive effect is reliably detected.

### 5. Conclusions

The described here non-linear feature extraction FASM algorithm is based on a model of auditory suppression. It is perceived as suitable mean to robustify ASR because allows to increase the feature invariance to noise, particularly the white noise, which was used in the current experiment. The measured improvement in the feature invariance is achieved at the cost of a just minor informativeness loss. The main achievement, reported in the present paper, in the fact, that measured robustness advantages are achieved with the help of the reduced complexity suppression model, which adds only a minor extra computational effort, compared to the very effective baseline WDFT-based computation of the perceptual spectrum. The proposed FASM algorithm is way more computationally efficient compared to the traditional auditory suppression modelling [1,2,9].

### 6. References


